PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference GB020049	FOR FURTHER ACTION	See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/GB 03/03745	International filing date *(day/month/year)* 29.08.2003	Priority date *(day/month/year)* 07.09.2002
International Patent Classification (IPC) or both national classification and IPC H04L29/06		
Applicant INTERNATIONAL BUSINESS MACHINES CORPORATION et al.		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 6 sheets, including this cover sheet.

   ☒ This report is also accompanied by ANNEXES, i.e. sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

   These annexes consist of a total of 9 sheets.

3. This report contains indications relating to the following items:

   - I ☒ Basis of the opinion
   - II ☐ Priority
   - III ☐ Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
   - IV ☐ Lack of unity of invention
   - V ☒ Reasoned statement under Rule 66.2(a)(ii) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
   - VI ☐ Certain documents cited
   - VII ☐ Certain defects in the international application
   - VIII ☐ Certain observations on the international application

Date of submission of the demand	Date of completion of this report
22.10.2003	24.08.2004
Name and mailing address of the international preliminary examining authority: European Patent Office D-80298 Munich Tel. +49 89 2399 - 0 Tx: 523656 epmu d Fax: +49 89 2399 - 4465	Authorized Officer Lopez Monclus, I. Telephone No. +49 89 2399-7113

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No. PCT/GB 03/03745

I. Basis of the report

1. With regard to the **elements** of the international application *(Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rules 70.16 and 70.17))*:

Description, Pages

5-9	as originally filed
1-4, 4a	received on 13.04.2004 with letter of 08.04.2004

Claims, Numbers

1-21	received on 13.04.2004 with letter of 08.04.2004

Drawings, Sheets

1/7-7/7	as originally filed

2. With regard to the **language**, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.

   These elements were available or furnished to this Authority in the following language: , which is:

   ☐ the language of a translation furnished for the purposes of the international search (under Rule 23.1(b)).

   ☐ the language of publication of the international application (under Rule 48.3(b)).

   ☐ the language of a translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

   ☐ contained in the international application in written form.

   ☐ filed together with the international application in computer readable form.

   ☐ furnished subsequently to this Authority in written form.

   ☐ furnished subsequently to this Authority in computer readable form.

   ☐ The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

   ☐ The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. The amendments have resulted in the cancellation of:

   ☐ the description, pages:

   ☐ the claims, Nos.:

   ☐ the drawings, sheets:

5. ☐ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed (Rule 70.2(c)).

   (Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.)

6. Additional observations, if necessary:

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty (N)	Yes:	Claims	1-21
	No:	Claims	
Inventive step (IS)	Yes:	Claims	1-21
	No:	Claims	
Industrial applicability (IA)	Yes:	Claims	1-21
	No:	Claims	

2. Citations and explanations

   see separate sheet

Re Item V

Reasoned statement with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

I

Reference is made to the following documents:

D1: US 2002/002611 A1 (VANGE MARK) 3 January 2002 (2002-01-03)
D2: WO 00/62502 A (FAN CHENGGONG ;LEMAHIEU PAUL (US); LOVE PHILIP (US); BRUCK JEHOSHU) 19 October 2000 (2000-10-19)

II

The subject-matter of the present application relates to a method (independent claim 1), a system (independent claim 11) and a computer program (independent claim 21) for the remote and dynamic configuration of a server to facilitate capacity on demand.

The closest prior art is document D1 which describes a system for shifting functionality between multiple web servers in a coordinated multi-server site. In order to improve conventional Internet service models, it adds a series of front-end servers, where content and functionality of the originating web servers can be transferred, closer to the user. A management component is disclosed which dynamically maintains and shifts content and functionality between originating servers and the various front end servers. D1 therefore discloses an accelerated caching service which shifts content by switching URLs to available front-end machines, which are configured in advance for that purpose.

The problem with the system described in D1 is that, since these front end servers are preconfigured and if the performance of one or more of them degradates, no more resources can be added. The problem to be solved by the present invention may be regarded as how to optimise the performance of a server system by being able to allocate or de-allocate server resources in response to a degradation of the performance.

In order to do so, the present application performs dynamic configuration is in response to collating performance data from a plurality of servers. A control server analyses said performance data and identifies if one or more of them have reached a predetermined threshold. If this is the case, it adjusts the allocation of the resources of the servers by issuing a configuration update to a dynamic configuration module of said servers.

The advantage of the present invention is that it allows the reconfiguration of the system without having to reconfigure and restart the servers, and allows for the optimisation of the system by allocating and deallocating server resources "on the fly" depending on the actual server's performance.

D2 is directed to the field of load balancing and does not teach or suggest the dynamic configuration of server resources in response to the performance statistics and connection settings of a particular server

An inventive activity is recognised and claims 1, 11 and 21 therefore fulfill the requirements of Article 33(3) PCT. Dependent claims 2-10 and 12-20 are also inventive by virtue of their dependency on claims 1 and 11, and therefore fulfill also the requirements of Article 33(3) PCT.

Certain defects in the international application (form or content).

1. Independent claim 11 is not properly cast in the two-part form (Rule 6.3(b) PCT), with those features which in combination are part of the prior art (see document D1) being placed in the preamble.

2. Reference sign 140 is missing in claim 1, line 12: "...to a second server;" Rule 6.2(b) PCT.

3. Claim 1 is not clear due to what seem to be clerical errors:

"...update instruction for the first server (140) or the third server (145) ..." should presumably read "...update instruction for the first server (145) or the third server (150)..." ;

4. Claims 1 is not clear because of the incorrect use of the parentheses made when referencing steps (a), (b), (c), (d) and (e).

According to Rule 6.2(b) references to the drawings should only be placed in parentheses after the corresponding features. This is however not the case of the terms in parentheses in claim 1.

Certain observations on the international application (clarity).

1. In claim 1 it is now not clear the actions taken by the second server:

" (140) analysing the received performance data..." should presumably read "analysing by the second server (140) the received performance data collated in step c."

" (140) adjusting the allocation of the resources..." should presumably read "adjusting by the second server (140) the allocation of the resources"

REMOTE DYNAMIC CONFIGURATION OF A WEB SERVER TO FACILITATE CAPACITY ON DEMAND

Field of the Invention

The invention relates to the field of network services and in particular to the remote and dynamic configuration of a server to provide server capacity on demand.

Background of the Invention

Many companies offer hosting services to provide customers with a secure; robust and flexible infrastructure in which to host a variety of applications for example web applications such as on-line banking, on-line shopping, information services and hosting service such as 'pay for used capacity' which, enable a customer to only pay for the processing power that they use and allows the customer to deploy the most up-to-date equipment in a cost effective manner. Hosting services provide many businesses with an alternative solution to building and running their technology infrastructure in-house by tapping into computer systems in other company's data centers to provide the management of software applications and hardware resources such as servers. The management and administration of these servers and services provide a tremendous challenge to many hosting companies, as a key problem with the management and administration of the servers within an environment such as a server farm is the rigid infrastructure and architecture of the servers due to the definition of roles the servers play in relation to the data the servers are publishing.

The rigid allocation of a resource to for example, a web server which supports a particular customer's product can result in an under utilized yet expensive web server resource not being used to it's full capability, while other web servers supporting other products are stretched to the point of breaking. Further today's current state of the art web server software is complex and flexible and can be configured and extended through Application Protocol Interfaces (API) to facilitate powerful processing beyond the standard serving of basic Hypertext Markup Language (HTML) pages. However current deployment practices and technologies require that the web servers are configured manually using either configuration files or binary registries and thus remain static during their operation, handling requests for a specific website or websites. When an additional or a different resource for a particular HTML page or a

different web site URL needs publishing or a need arises for a new server to be added to, or removed from a server pool, the servers must be restarted for any change to take effect. This requires a tremendous administration effort on the part of the server farm administrator because it takes a considerable amount of time to manually configure a server and it is not always convenient to shut down a server and restart the server to allow for any changes to take effect as this can cause loss of service for a period of time.

US patent application publication US 2002/0002602 (2602) describes a system for serving web pages to a client in response to a client request specifying a resource that aims to serve a web page in a coordinated fashion from multiple cooperating web servers and maintaining a reliable connection so that the server and clients remain synchronized and information is not lost. In order for the above to take place such that a web server processes a URL and associates the URL with a data source, the web server will require manual intervention and the web server will have to be shut down and restarted for the changes to take effect.

US patent application, publication number US 2002/002611 A1 discloses a system for providing network functionality from a plurality of network connected servers to at least one network connected client computer. US 2002/002611 is concerned with providing an accelerated proxy/caching service to available pre-configured front-end servers.

WO 00/62502 discloses a distributed server cluster for controlling network traffic. WO 00/62502 is directed towards a load balancing system that on detection of a failed server automatically shifts the network traffic from the failed machine to one or more operational machines.

Disclosure of the Invention

Viewed from a first aspect the present invention provides a method for the remote and dynamic configuration of a server to provide server capacity on demand, the method comprising the steps of: (a) receiving a request for a resource by a first server (145) from a client device (100); characterised in that (b) routing the client request for the resource to a dynamic content module (215), the dynamic content module (215) identifying an available third server (150) from which the requested resource can be served and routing the requested resource to the client device (100); (c) collating performance data from the first (145) and third server (150) and reporting the performance data to a second server; (d) analysing the

performance data collated in step (c) to determine performance capabilities of the first and the third server (150) and identifying if the first or the third server has reached a predetermined threshold; and (e) adjusting the allocation of the resources of first server (145) or the third server (150) in response to step (d) and issuing a configuration update instruction for the first (140) server or the third server (145) to a dynamic configuration module (216) of the first server (145) and determining if a resource update is successful.

The present invention advantageously allows for the optimisation of a servers performance such that a server can be allocated or deallocated from a server pool depending on the server's performance for example if a server is stretched to capacity or equally if a server is under utilised. Further using modular autonomic computing components the system is able to configure and reconfigure itself under varying and unpredictable conditions. Further the method allows for making a change to a hardware resource or a data resource without the need for the server to be restarted as with other prior art systems.

Preferably the invention provides for the dynamic content module requesting a connection configuration file for the third server from the dynamic configuration module. This allows for the first server to determine which data source server has the resource requested by the client.

Preferably the invention provides for the adjusting of the allocation or deallocation of one or both the first server and the third server from a free server pool or to a free server resource pool. The present invention can therefore monitor a servers performance and either add an additional resource or de-allocate a resource depending on the work load of the server.

Preferably the invention provides for the requested resource being decoupled from the first server allowing the introduction of a new service or the removal of a redundant service.

Viewed from another aspect the present invention provides a system for the remote and dynamic configuration of a server to provide capacity on demand, the system comprising a client device for requesting and receiving a resource in a communications network, the system further comprising: a first server, the first server comprising a dynamic content module, a dynamic configuration module and a reporting module; means for routing the

client request for the resource to the dynamic content module; means for the dynamic content module identifying a third server from which the requested resource can be served and means for retrieving a connection configuration file associated with the third server stored in the dynamic configuration module; means for the reporting module collating performance data from the first server and the third server and means for routing the performance data to a second server; a second server comprising an analyser module, a resource allocation module and a resource update module, the second server sending the performance data to the analyser module; means for the analyser module determining the performance capabilities of the first server and the third server and means for identifying if the first server and third server has reached a predetermined threshold; means for the resource allocation module adjusting the allocation of one or both the first server and the third server in response to the identifying means; means for the resource update module issuing an configuration update instruction for one or both the first server and the third server to the dynamic configuration module of the first server and means for determining if a resource update is successful; and a third server comprising one or more resources and means for the third server to serve a requested resource to the first server.

Another advantage is that an HTTP URL can be processed by a server and associated with a remote data source without the need for the server to be restarted or requiring manual intervention. A further advantage of the present invention is the secure and centralised administration for the dynamic plug and play of data sources and hardware resources such that a web site can be served 'on the fly'. Another advantage of the present invention is for the provision of multiple protocol support for other mechanisms such as FTP, XML, SOAP and file sharing.

Brief description of the drawings

The invention will now be described by way of example only, with reference to the accompanying drawings, in which:

Figure 1 illustrates a server farm in which the present invention may be implemented in accordance with a preferred embodiment of the present invention;

Figure 2 illustrates a block diagram detailing an overview of the components of the system in accordance with a preferred embodiment of the present invention;

Figure 3 illustrates a flowchart detailing the function of the dynamic content module of the web server of Figure 2 in accordance with a preferred embodiment of the present invention;

Figure 4 illustrates a flowchart detailing the function of the reporting module of the web server of Figure 2 in accordance with a preferred embodiment of the present invention;

CLAIMS

1. A method for the remote and dynamic configuration of a server (145, 150,155) to provide server capacity on demand the method comprising the steps of:

(a) receiving a request for a resource by a first server (145) from a client device (100); characterised in that

(b) routing the client request for the resource to a dynamic content module (215), the dynamic content module (215) identifying an available third server (150) from which the requested resource can be served and routing the requested resource to the client device (100);

(c) collating performance data from the first (145) and third server (150) and the first server reporting the performance data to a second server;

(d) (140) analysing the received performance data collated in step (c) to determine the performance capabilities of the first and the third server (150) and identifying if the first or the third server has reached a predetermined threshold; and

(e) (140) adjusting the allocation of the resources of the first server (145) or the third server (150) in response to step (d) and issuing a configuration update instruction for the first server (140) or the third server (145) to a dynamic configuration module (216) of the first server (145) and determining if a resource update is successful.

2. A method as claimed in claim 1 wherein the dynamic content module (215) further comprises requesting a connection configuration file for the third server from the dynamic configuration module (216).

3. A method as claimed in claim 1 wherein the dynamic configuration module (216) stores configuration settings for one or both the first server (145) and the third server (150).

4. A method as claimed in claim 1, wherein the step of adjusting the allocation of one or both the first server (145) and the third server (150) further comprises allocating an additional server from a free server resource pool.

5. A method as claimed in claim 1, wherein the step of adjusting the allocation of one or both the first server (145) and the third server (150) further comprises de-allocating the first server (145) or the third server (150) from an allocated resource pool to a free server resource pool.

6. A method as claimed in claim 1 or 2 wherein the first server (145) and the second server (140) communicate with each other through XML data streams.

7. A method as claimed in claim 1 wherein the second server (140) is a management server providing a central control point for one or more first servers (145) and one or more third servers (150).

8. A method as claimed in claim 1 wherein the requested resource is decoupled from the first server (140) allowing the introduction of a new service or the removal of a redundant service.

9. A method as claimed in claim 1 wherein the first server (145) is plurality of servers.

10. A method as claimed in claim 1 wherein the third server (150) is plurality of servers.

11. A system for the remote and dynamic configuration of a server to provide server capacity on demand, the system being for use with a client device which requests and receives a resource in a communications network, the system comprising:

a first server (145), the first server (145) comprising a dynamic content module (215), a dynamic configuration module (216) and a reporting module (205);

the first server (145) further comprising:

means for routing the client request for the resource to the dynamic content module (215);

means for the dynamic content module (215) identifying a third server (150) from which the requested resource can be served and means for retrieving a configuration file associated with the third server (150) stored in the dynamic configuration module; and

means for the reporting module (205) collating performance data from the first server (145) and the third server (150) and means for routing the performance data to a second server (140);

a second server comprising an analyser module (230), a resource allocation module (235) and a resource update module (240), the second server (140) sending the performance data to the analyser module (230); the second server (140) further comprising:

means for the analyser module (230) determining the performance capabilities of the first server (145) and the third (150) server and means for identifying if the first server (145) and the third server (150) has reached a predetermined threshold;

means for the resource allocation module (235) adjusting the allocation of one or both the first server (145) and the third server (150) in response to the identifying means; and

means for the resource update module (240) issuing a configuration update instruction for one or both the first server (145) and the third server (150) to the dynamic configuration module (216) of the first server (145) and means for determining if the allocation or de-allocation of one or both the first server (145) and the third server (150) has been successful; and

a third server (150) comprising one or more resources and means for the third server (150) to serve a requested resource to the first server (145).

12. A system as claimed in claim 11 wherein means for adjusting the allocation of one or both the first server (145) and the third server (150) further comprises means for allocating an additional server from a free server resource pool.

13. A system as claimed in claim 11 wherein means for adjusting the resource allocation of the first server (145) or the third server (150) further comprises means for de-allocating one or both the first server (145) and the third server (150) from an allocated resource pool to a free server resource pool.

14. A system as claimed in claim 11 wherein means for communication between the first server (145) and the second server (140) is through XML data streams.

15. A system as claimed in claim 11 wherein means for the connection configuration file comprises connection settings for the first (145) and the third server (150).

16. A system as claimed in claim 11 wherein the dynamic configuration module (216) provides means for storing configuration settings for a first server (145) and a third server (150).

17. A system as claimed in claim 11 wherein the second server (140) is a management server providing means for a central control point for the first and the third server (150).

18. A system as claimed in claim 11 wherein means for the requested resource is decoupled from the first server (145) allowing means for the introduction of a new service or the removal of a redundant service.

19. A system as claimed in claim 11 wherein the first server (145) is plurality of servers.

20. A system as claimed in claim 11 wherein the third server (150) is plurality of servers.

21. A computer program product comprising computer program code stored on a computer readable storage medium, which when executed on a data processing system, instructs the data processing system to carry out the method as claimed in claim 1.